1 Part 1: Contextualizing the Data

Let's try to understand the background of our dataset before diving into a full-scale analysis.

1.1 Question 1

1.1.1 Part 1

Based on the columns present in this data set and the values that they take, what do you think each row represents? That is, what is the granularity of this data set?

Each row represents a housing sale record. It means the granularity of this data set is a single sale of a property.

1.1.2 Part 2

Why do you think this data was collected? For what purposes? By whom?

This question calls for your speculation and is looking for thoughtfulness, not correctness.

I think this data was collected to predict or evaluate housing sale prices in Cook County. It was most likely collected by real estate companies.

1.1.3 Part 3

Certain variables in this data set contain information that either directly contains demographic information (data on people) or could when linked to other data sets. Identify at least one demographic-related variable and explain the nature of the demographic data it embeds.

I think the variable "Neighborhood Code" is a demographic-related variable. Because housing sale prices can vary widely in different neighborhoods. So it could tell people's income level in those locations.

1.1.4 Part 4

Craft at least two questions about housing in Cook County that can be answered with this data set and provide the type of analytical tool you would use to answer it (e.g. "I would create a _____ plot of _____ and ***" or "I would calculate the*** [summary statistic] for _____ and _____"). Be sure to reference the columns that you would use and any additional data sets you would need to answer that question.

- Question 1. What are the differences in average housing sale price in different neighborhoods?

I would create a bar plot of "Neighborhood Code" and mean "Sale Price".

- Question 2. What is the distribution of sale prices in Cook County?

I would create a boxplot of "Sale Price".

1.2 Question 2

1.2.1 Part 1

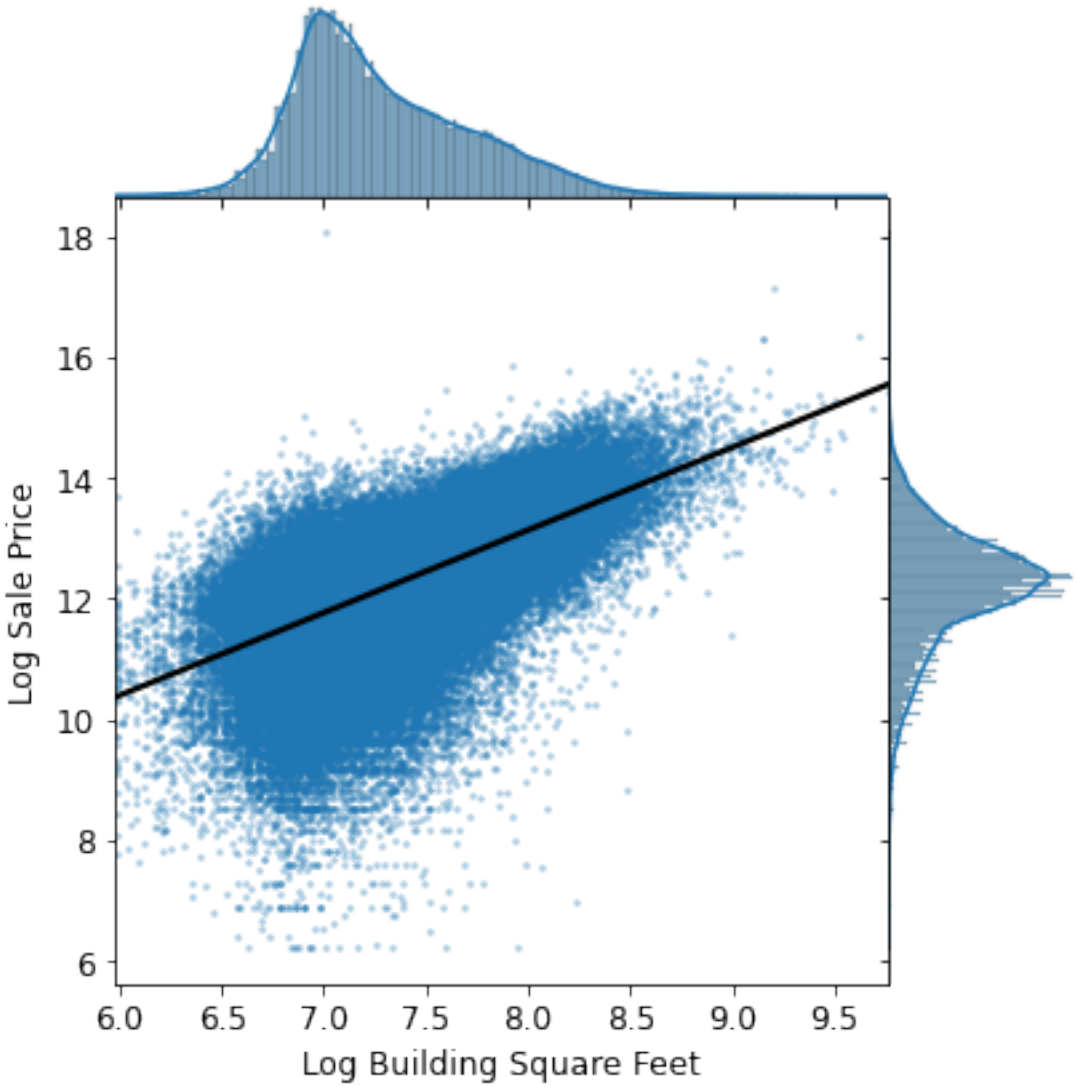
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards as the autograder may not work otherwise.

Too many sale prices are $1 so that the visualizations became not clear. I think we should only keep sale prices that over \$1 or another specific threshold like \$500.

1.2.2 Part 3

As shown below, we created a joint plot with `Log Building Square Feet` on the x-axis, and `Log Sale Price` on the y-axis. In addition, we fit a simple linear regression line through the bivariate scatter plot in the middle.

Based on the following plot, does there exist a correlation between `Log Sale Price` and `Log Building Square Feet`? Would `Log Building Square Feet` make a good candidate as one of the features for our model?



Yes, there exist a positive correlation between Log Sale Price and Log Building Square Feet. Thus, Log

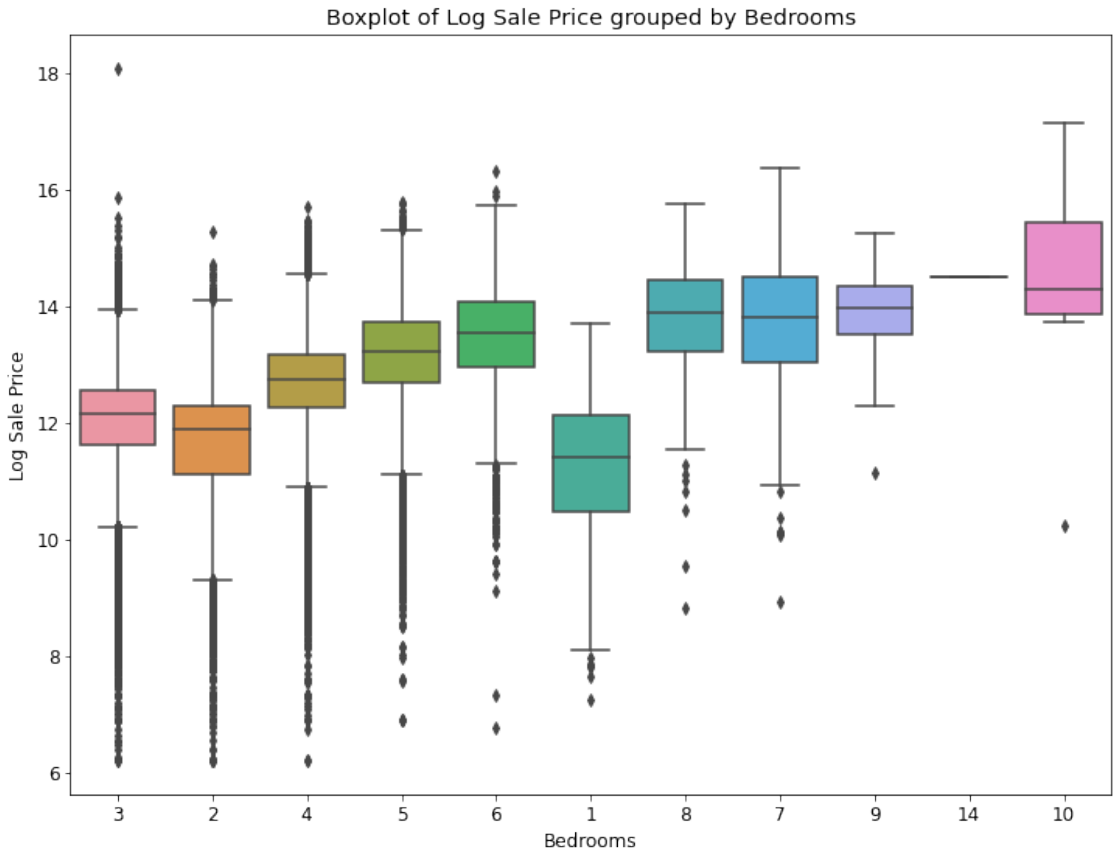
Building Square Feet should be a good candidate as one of the features for our linear regression model.

1.2.3 Part 3

Create a visualization that clearly and succintly shows if there exists an association between `Bedrooms` and `Log Sale Price`. A good visualization should satisfy the following requirements: - It should avoid overplotting. - It should have clearly labeled axes and succinct title. - It should convey the strength of the correlation between the sale price and the number of rooms.

Hint: A direct scatter plot of the sale price against the number of rooms for all of the households in our training data might risk overplotting.

```
In [25]: sns.boxplot(x='Bedrooms', y='Log Sale Price', data=training_data)
         plt.xlabel('Bedrooms')
         plt.ylabel('Log Sale Price')
         plt.title('Boxplot of Log Sale Price grouped by Bedrooms')
         plt.show()
```



1.2.4 Part 3

It looks a lot better now than before, right? Based on the plot above, what can be said about the relationship between the houses' `Log Sale Price` and their neighborhoods?

Yes, the zoomed in plot is better than before. Based on the plot, I don't think Log Sale Price and neighborhood have correlation.